DOCUMENT RESUME

ED 259 005 TM 850 374

AUTHOR O'Brien, Michael; Hampilos, John P.
TITLE The Feasibility of Creating an Item Bank from a Teacher-Made Test, Using the Rasch Model.
PUB DATE 24 Apr 84
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 24, 1984).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Assisted Testing; Criterion Referenced Tests; Difficulty Level; Higher Education; *Item Analysis; *Item Banks; *Latent Trait Theory; *Teacher Made Tests; *Test Construction; Test Items; Test Reliability; Test Validity
IDENTIFIERS *Rasch Model

ABSTRACT

The feasibility of creating an item bank from a teacher-made test was examined in two comparable sections of a graduate-level introductory measurement course. The 67-item midterm examination contained multiple-choice and master matching items, which required higher level cognitive processes such as application and analysis. The feasibility of banking the test items was determined by the comparability of invariant item difficulty (item reliability) and item quality (item validity). The test was administered to Spring and Fall classes, and a Rasch latent trait analysis was performed on each test administration. The reliability of each item difficulty was determined by comparing the difference between Rasch difficulty estimates from each class for each item against their calibration error. Item validity was studied by examining the extent to which each item approximated an a priori rating used for criterion-referenced grading. Test reliability and test validity results indicated that the majority of items could be usefully banked. In the future, new test items could be created using the same procedures. The development of computer software to facilitate item analysis was recommended. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED259005

THE FEASIBILITY OF CREATING AN ITEM BANK FROM A TEACHER-MADE TEST, USING THE RASCH MODEL

MICHAEL L. O'BRIEN AND JOHN P. HAMPILOS

TEACHERS COLLEGE, COLUMBIA UNIVERSITY

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

O'Brien, Michael L.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April 24, 1984.

TM 850 374

ABSTRACT

This study focused on the feasibility of creating an item bank for a teacher-made test. Procedures for investigating reliability and validity at the level of the individual item are illustrated. The results indicated that the prospect of creating useful item banks for teacher-made tests would be favorable, given consistent instructional techniques and careful attention to item construction. In addition, a procedure for establishing criterion-referenced grading is discussed.

Historically, teacher-made tests provided a basis for grading student achievement. However, these measures suffered from two related constraints. First, the teacher felt obligated to re-use the same "good test (i.e. set of items)" because the set of items were of known, reliable difficulty. Second, the test included a set of items deemed to provide relevant, valid information for assigning each student a grade. These constraints logically depend upon proper understanding of the instructional conditions that make an item reliable and valid. Practically, it is clearly unwise to re-use a set of items for fear of test security. However, without some procedure to guarantee the "success" of a new configuration of items, test reliability and validity would seemingly be left to chance. Item banking provides a solution to these concerns by focusing on reliability and validity at the level of the individual item. Latent trait technology provides a psychometric method to enhance the feasibility of developing reliable and valid tests from individually calibrated and validated items. Further, recent developments in criterion-referenced testing facilitates a practical procedure for assigning grades without pitting students against each other in competition for grades. Therefore, combining item banking, latent trait procedures, and criterion-referenced grading should provide a solution to the problem of assembling "new," reliable and valid teacher-made tests.

THEORETICAL FRAMEWORK

Ironically, while theory leading to the item banking of teacher-made tests has been established, it has yet to be put into practice! In part, the problem is the lack of required technology available to teachers. However, the more serious problem is the need to synthesize and to demonstrate the feasibility of theories relevant to assembling reliable and valid tests from item banks calibrated from applying latent trait technology.

Choppin (1969) conducted a landmark study of the feasibility of banking achievement test items. By having teachers write achievement test items to conform to a blueprint (Thorndike and Hagen, 1955), he investigated whether the items would be of sufficient reliability and validity. Using Rasch latent trait item analysis, he demonstrated the utility of sample-free item calibration for item banking. If a given item measured the same, underlying item parameter across successive samples of students, using the Rasch model would facilitate the calibration of a bank of items of known, reliable difficulty. Further, the item must demonstrate fit to the Rasch model in order to provide evidence of unidimensionality. That is, does the item measure the common, underlying latent trait? If individual items can be calibrated reliably along a single, underlying latent trait, then the selection of any given set of items conforming to blueprint specifications can be selected from the item bank. The resulting test would, by definition, be expected to be of high reliability (i.e. internal consistency and stable difficulty).

Choppin's work clearly set the stage for solving the age-old problem of classroom measurement procedures. However, he found that instructional effects put the validity of the item bank in question. Because teachers from different schools perceived the content and process of a given item differently, item banking might endanger content validity. Thus, the reliability of an item would, theoretically, be related to its validity. This problem underscored the need to further consider item validity before banking a given item shown to be reliable. Loevinger (1958) and Wright and Stone (1979) discussed this problem extensively and concluded that the test constructor should make an a priori hypothesis about the expected difficulty of each test item. Thus, when test experts convene to establish content validity for each item, they should rate each item for its expected level of difficulty.

Item validity, therefore, could be investigated by comparing the observed v. expected difficulty of each item. The item bank would, therefore, contain items shown to evidence reliable and "valid" item difficulty.

The need to posit hypotheses about teacher-made tests was underscored by Ebel (1983). He argued that, practically speaking, teachers seldom attempt to validate tests formally. However, if an item bank were established, test validity would be attainable from any given set of items across new samples of students. Further, criterion-referenced grading (Berk, 1980) could be validated by relating item difficulty to how a typical, say "A-" student should perform on that item. That is, what is the probability that an A- student will answer Item I correctly? Nedelsky (1954) and Glass (1978) discussed this type of procedure for determining criterion scores but failed to recognize the need to validate the difficulty of each item. Having established the reliability and validity of each item, assigning a criterion-referenced grade would depend only on the teacher's ability to identify the characteristics needed to achieve each possible grade. That is, what is the probability that a student with "latent A- ability" would answer correctly an item with a known difficulty?

Thus, the two critical problems of assembling teacher-made tests can be theoretically solved by item banking. First, having a reliable item difficulty estimate for each item would facilitate the assembling of an equivalent set of items that would provide optimal information across ability level (Wright and Stone, 1979). Second, having a valid a priori rating of each item difficulty would facilitate criterion-referenced grading for any given set of items (see Methods section).

PROBLEM

Item banking teacher-made test items would be useful for assembling better tests. However, the feasibility of creating item banks for this purpose

has yet to be demonstrated. Moreover, the procedures for constructing a useful item bank have not been previously established in a single study. By combining the theoretical contributions of previous research, it is possible to determine if teacher-made tests could be assembled from the item bank. Thus, the central research question posited for this study focused on the feasibility of establishing sufficient item reliability and item validity to facilitate banking items for teacher-made tests. As part of this central question, how might item reliability and item validity be related?

METHODS

Test Development

The present study was based on the feasibility of creating an item bank for a midterm examination for a graduate, introdutory measurement course at Teachers College, Columbia University. The course covered introductory psychometric concepts, including Use of Tests, Test Construction, and Statistical Analysis. The textbook used for the course was Thorndike and Hagen's Measurement and Evaluation in Psychology and Education, 4th Edition, Wiley, 1977.

The feasibility of creating an item bank for this course was based upon administering the same, 67-item midterm examination to two graduate Introduction to Measurement classes during Fall Semester, 1983 and Spring Semester, 1984. The course was taught by the same instructor who used the same set of lecture notes for each class. Each class contained approximately 60 students majoring in Psychology, Education, or Nursing. The majority of students were female. Additional analyses showed that there were no differences in achievement as a function of major or sex.

The 67-item midterm examination was designed by the instructor to tap

appropriate content and process. The items were of multiple choice or master matching format. Since the students were allowed to bring in three pages of notes to the exam, the test focused primarily on higher level cognitive processes, such as Application and Analysis. The Bloom taxonomy was used to categorize each item in terms of content and process (Bloom, Hastings, and Madaus, 1971). Content validity was established by a review of the test among students enrolled in an advanced psychometrics course at Teachers College.

Each item was also rated for expected item difficulty by a group of advanced, doctoral measurement students and the course instructor. The item rating, referred to as "C_I," was based on a scale of 1 to 10, were "1" would indicate a hard item. The ratings were referenced to levels of ability being measured by each item. Figure 1 shows the relationship between item rating and level of ability, as defined by letter grade.

Letter Grade :	A+	A	A-	B+	B	B-	C+	C	C-	F
Item Rating (C_I):	1	2	3	4	5	6	7	8	9	10

Figure 1.--Illustration of the referencing of letter grade to item rating.

In order to assign a rating for each item, the rater was asked to identify the point (i.e. letter grade) at which a student would achieve mastery of the skill being measured by the item. That is, at which level of ability is the item discriminating optimally between mastery and non-mastery? Mastery was operationally defined as an 80 percent probability of answering the item correctly. For example, the item shown in Figure 2 was assigned an item rating (C_I) of "3," indicating that this item should optimally measure mastery attained by a student of a latent "A-" ability level. That is, Item 55 was judged to be somewhat difficult and, therefore, appropriate for optimally measuring an "A- student."

Item 55. In the normal distribution, if the $\overline{X}$ = 50 and SD = 10, about how many cases in 100 should fall between 40 and 60?

A. 34
B. 68
C. 95
D. cannot be estimated

Figure 2.--Sample item from the midterm examination.

The probability of success on a given item could now be referenced <u>across</u> possible letter grades. Probability of a correct response could be estimated for each item as a function of C_I and number of options, as shown in Figure 3.

Relation between ability and difficulty	Probable success on 4-option item	Probable success on 5-option item
C_I + > 2	100	100
C_I + 2	90	90
C_I + 1	90	90
C_I	80	80
C_I - 1	75	80
C_I - 2	50	60
C_I - 3	50	40
C_I - 4	25	20
C_I - > 4	25	20

Figure 3.--Illustration of probable success for each possible combination of person ability and item difficulty.

From Figure 3, it can be seen that the expected success probability can be set systematically for an item of any given difficulty, assuming that an examinee will successfully eliminate incorrect options as a function of his or her ability level. Figure 4 shows the expected success probabilities for Item 55.

Letter Grade:	A+	A	A-	B+	B	B-	C+	C	C-	F
C_I :	1	2	3*	4	5	6	7	8	9	10
Probable Success on Item 55 (4-options)	90	90	80	75	50	50	25	25	25	25

Figure 4.--Illustration of success probabilities on Item 55 (C_I = 3, 4-options).

The success probabilities for each of the 67 items are reported in Table 4. In order to determine a criterion score for the total test score, one would simply sum the success probabilities for each item at each possible letter grade and subtract one SE_M. Thus, for any given set of items in the item bank, the teacher could assemble an "optimal" overall test difficulty and have the criterion scores for each letter grade determined in advance! Inspection of Table 4 would facilitate better comprehension of the details of this procedure.

Test Administration

Each class took the test during a two-hour period. Students were given a test booklet and indicated their response to each item on an answer sheet. Test booklets and answer sheets were numbered to maintain test security.

Item Analyses

Item difficulty was estimated by applying the BICAL computer program for Rasch analysis of items scored dichotomously. A fit statistic, Fit-T, was also estimated from BICAL (Wright and Mead, 1976) to test the fit of each item to the Rasch model. The SPSS computer program was also applied to facilitate classical estimates for item difficulty (i.e. p-values) and total test internal consistency reliability (i.e. $KR\text{-}_{20}$). A Phi-coefficient (ϕ) could then be estimated to test the relationship between items accepted v. rejected on the basis of item reliability and item validity.

RESULTS AND DISCUSSION

Item Reliability

In order to investigate the reliability of each item, the Rasch item difficulty for each item was compared across the Fall v. Spring administration of each item. The results of this analysis are reported in Table 1. The technique for investigating reliability by this method are given by Wright and Stone (1979).

The overall correlation between the item difficulties from the two administrations was $r = 0.78$ ($p < .01$). The $KR\text{-}_{20}$ estimate from the Fall sample was 0.87; for Spring, 0.86 (the Spring class had slightly higher $\overline{X}$ and lower SD compared to the Fall sample). Thus, in general, the items evidenced highly stable difficulty across time and could, therefore, be banked. Individual tests of item reliability were performed by computing the ratio of the differences between the two Rasch difficulty estimates to th pooled standard error of calibration ($t = diff_{F-S} / Se_p$). The resulting "t-test" of item reliability identified individual items that were relatively harder in the Spring ($t < -2$) or that were relatively harder in the Fall ($t > 2$). Inspection of these items revealed that, for about 25 percent of the items, specific instructional effects could logically explain the relative shifting of the difficulty of an item. The implication of this finding is that more attention should be paid to standardizing clear instructional procedures!

Item fit to the Rasch model was also hypothesized to tap the consistency of each item with the underlying trait being measured by the test (i.e. achievement in measurement concepts). Table 2 presents the results of this investigation. First, comparing within Fall or Spring administration, it was clear that practically all of the items fit the Rasch model. However, the item with the worst misfit to the Rasch model in the Fall, Item 48 (Fit-t = 3.4), suffered from faulty item writing. It was revised for the Spring and concurrently

Table 1

Results of Item Reliability Analysis
and Final Item Bank Estimates

Item No.	d_F	$Se(d_F)$	d_S	$Se(d_S)$	$diff_{F-S}$	Sep	t	Bank Value d_B	$Se(d_B)$	C_I'	C/P[1]
01	-1.08	0.48	-0.31	0.45	-0.77	0.46	-1.7	-0.70	0.46	9	U/K
02	2.19	0.28	2.09	0.29	0.10	0.28	0.4	2.14	0.28	3	U/C
03	-0.12	0.35	-0.53	0.49	0.41	0.42	1.0	-0.34	0.42	5	U/K
04	-0.01	0.34	-1.55	0.74	1.54	0.54	2.9	-0.78	0.54	5	U/K
05	1.07	0.28	0.83	0.33	0.24	0.30	0.8	0.95	0.30	3	U/C
06	-0.87	0.44	-1.11	0.61	0.14	0.52	0.3	-0.99	0.52	5	U/C
07	0.21	0.32	0.05	0.40	0.16	0.36	0.4	0.13	0.36	3	U/Ap
08	-0.87	0.44	-0.78	0.54	-0.09	0.49	-0.2	-0.82	0.49	5	U/Ap
09	0.10	0.33	-0.53	0.49	0.63	0.41	1.5	-0.22	0.41	4	U/Ap
10	1.52	0.27	1.13	0.31	0.39	0.29	1.3	1.32	0.29	4	U/Ap
11	1.00	0.28	1.23	0.31	-0.23	0.30	-0.8	1.12	0.30	2	U/Ap
12	0.31	0.31	0.05	0.40	0.26	0.36	0.7	0.18	0.36	2	U/Ap
13	0.67	0.29	0.48	0.36	0.19	0.32	0.6	0.58	0.32	2	U/Ap
14	-0.87	0.44	-0.78	0.54	-0.09	0.49	-0.2	-0.82	0.49	7	C/K
15	-0.12	0.35	-0.78	0.54	0.66	0.44	1.5	-0.45	0.44	5	C/Ap
16	-2.07	0.73	-1.11	0.61	-0.96	0.67	-1.4	-1.59	0.67	6	C/K
17	0.31	0.31	0.34	0.37	-0.03	0.34	-0.1	0.32	0.34	4	C/An
18	0.10	0.33	-1.11	0.61	1.21	0.47	2.6	-0.50	0.47	6	C/Ap
19	0.40	0.31	0.05	0.40	0.35	0.36	1.0	0.22	0.36	2	C/An
20	-1.08	0.48	-1.11	0.61	0.03	0.54	0.1	-1.10	0.54	3	C/An
21	-0.87	0.44	-1.55	0.74	0.68	0.59	1.2	-1.21	0.59	5	C/Ap
22	-1.33	0.53	-1.11	0.61	-0.22	0.57	-0.4	-1.22	0.57	6	C/Ap
23	-2.07	0.73	-2.27	1.02	0.20	0.88	0.2	-2.17	0.88	9	C/Ap
24	-0.87	0.44	-0.53	0.49	-0.34	0.46	-0.7	-0.65	0.46	4	C/Ap
25	-0.38	0.38	-0.78	0.54	0.40	0.46	0.9	-0.50	0.46	5	C/Ap
26	-2.07	0.73	-2.27	1.02	0.20	0.88	0.2	-2.17	0.88	8	C/Ap
27	1.30	0.28	1.13	0.31	0.17	0.30	0.6	1.22	0.30	4	C/Ap
28	-0.12	0.35	1.13	0.31	-1.25	0.33	-3.8	0.62	0.33	4	C/Ap
29	0.75	0.29	1.59	0.30	-0.84	0.30	-2.8	1.17	0.30	3	C/Ap
30	0.31	0.31	-0.12	0.43	0.43	0.37	1.2	0.10	0.37	4	C/Ap
31	-0.01	0.34	0.34	0.37	-0.35	0.36	-1.0	0.16	0.36	6	C/Ap
32	-0.53	0.39	-2.27	1.02	1.74	0.70	2.5	-1.40	0.70	7	C/Ap
33	0.49	0.30	-0.12	0.43	0.61	0.36	1.7	0.18	0.36	6	C/Ap
34	-0.53	0.39	-1.11	0.61	0.58	0.50	1.2	-0.82	0.50	6	C/Ap
35	0.21	0.32	0.20	0.39	0.01	0.36	0.0	0.20	0.36	6	C/Ap
36	1.15	0.28	0.72	0.34	0.43	0.31	1.4	0.94	0.31	3	C/Ap
37	-2.79	1.02	-2.27	1.02	0.52	1.02	0.5	-2.53	1.02	8	C/Ap
38	-0.01	0.34	1.13	0.31	-1.14	0.32	-3.6	0.56	0.32	5	C/Ap
39	-2.07	0.73	-0.53	0.49	-1.54	0.61	-2.5	-1.30	0.61	6	C/Ap
40	1.07	0.28	1.13	0.31	-0.06	0.30	-0.2	1.10	0.30	4	C/Ap

TABLE 1, continued

Item No.	d_F	$Se(d_F)$	d_S	$Se(d_S)$	$diff_{F-S}$	Sep	t	Bank Value d_B	$Se(d_B)$	C_I'	C/P
41	-0.53	0.39	0.20	0.39	-0.73	0.39	-1.9	-0.16	0.39	6	C/Ap
42	-1.33	0.53	-1.11	0.61	-0.22	0.57	-0.4	-1.22	0.57	8	C/Ap
43	0.10	0.33	0.48	0.36	-0.38	0.34	-1.1	0.29	0.34	6	C/Ap
44	2.19	0.28	2.25	0.29	-0.06	0.28	-0.2	2.22	0.28	2	C/Ap
45	1.45	0.27	1.76	0.29	-0.31	0.28	-1.1	1.60	0.28	5	C/Ap
46	-1.64	0.60	-2.27	1.02	0.63	0.81	0.8	-1.96	0.81	9	C/Ap
47	0.67	0.29	1.03	0.32	-0.36	0.30	-1.2	0.85	0.30	4	C/Ap
48	1.82	0.27	0.93	0.32	0.89	0.30	3.0	1.38	0.30	1	C/Ap
49	-0.53	0.39	0.48	0.36	-1.01	0.38	-2.7	-0.02	0.38	8	S/K
50	-0.87	0.44	-1.55	0.74	0.68	0.59	1.2	-1.21	0.59	8	S/K
51	-0.01	0.34	0.20	0.39	-0.21	0.36	-0.6	0.10	0.36	4	S/C
52	0.31	0.31	-0.53	0.49	0.84	0.40	2.1	0.11	0.40	4	S/Ap
53	-0.87	0.44	-1.11	0.61	0.24	0.52	0.5	-0.99	0.52	6	S/Ap
54	0.21	0.32	0.05	0.40	0.16	0.36	0.4	0.13	0.36	6	S/K
55	0.49	0.30	0.60	0.35	-0.11	0.32	-0.3	0.54	0.32	3	S/Ap
56	1.30	0.28	1.67	0.29	-0.37	0.28	-1.3	1.48	0.28	2	S/Ap
57	-0.25	0.36	-0.31	0.45	0.06	0.40	0.2	-0.28	0.40	4	S/Ap
58	-0.01	0.34	1.32	0.30	-1.33	0.32	-4.2	0.66	0.32	3	S/K
59	2.12	0.28	0.72	0.34	1.40	0.31	4.5	1.42	0.31	4	S/K
60	1.52	0.27	0.48	0.36	1.04	0.32	3.2	1.00	0.32	5	S/Ap
61	0.49	0.30	0.20	0.39	0.29	0.34	0.9	0.35	0.34	4	S/Ap
62	2.12	0.28	1.67	0.29	0.45	0.28	1.6	1.85	0.28	3	S/K
63	0.40	0.31	0.93	0.32	-0.53	0.32	-1.7	0.67	0.32	4	S/C
64	-1.33	0.53	-0.31	0.45	-1.02	0.49	-2.1	-0.82	0.49	8	S/C
65	1.45	0.27	0.72	0.34	0.73	0.30	2.4	1.08	0.30	2	S/C
66	-0.53	0.39	-0.31	0.45	-0.22	0.42	-0.5	-0.42	0.42	4	S/Ap
67	-0.12	0.35	2.75	0.30	-2.87	0.32	-9.0	1.44	0.32	1	S/Ap

---

[1] C/P denotes Content/Process
Content Categories:
U denotes Test Use
C " Test Construction
S " Statistics
Process Categories:
K denotes Knowledge
C " Comprehension
Ap " Application
An " Analysis

Table 2

Summary of Rasch Item Fit Statistics

Item No.	Fit-t_F	Fit-t_S	t
01	0.4	-0.4	0.8
02	0.7	2.7	2.0
03	0.6	0.0	0.6
04	0.7	0.1	0.6
05	2.4	1.1	1.3
06	-0.3	-0.0	0.3
07	-0.2	-0.2	0.0
08	-0.6	-0.2	0.4
09	0.1	0.1	0.0
10	1.3	0.5	0.8
11	0.9	-0.8	1.7
12	-0.1	-0.3	0.2
13	0.4	0.7	0.3
14	0.2	-0.3	0.5
15	0.1	0.5	0.4
16	0.1	-0.0	0.1
17	0.0	1.2	1.2
18	-0.6	-0.2	0.4
19	-1.3	-0.2	1.1
20	-0.3	-0.0	0.3
21	-0.2	-0.0	0.2
22	0.0	-0.3	0.3
23	-0.1	0.3	0.4
24	0.1	0.1	0.0
25	-0.1	-0.2	0.1
26	-0.1	0.1	0.2
27	0.0	1.6	1.6
28	0.6	1.4	0.8
29	-0.8	1.1	1.9
30	-0.3	-0.9	0.6
31	-0.8	0.3	1.1
32	-0.8	0.2	1.0
33	-1.3	-0.8	0.5
34	0.0	0.2	0.2
35	0.0	0.4	0.4
36	0.4	-1.1	1.5
37	0.2	0.3	0.1
38	0.1	-0.1	0.2
39	0.2	0.4	0.2
40	-0.0	1.9	1.9
41	-0.6	-0.2	0.4
42	-0.0	-0.2	0.2
43	0.5	0.7	0.2
44	1.2	-0.2	1.4
45	0.7	-0.9	1.6
46	-0.1	0.2	0.3
47	1.0	-0.1	1.1
48	3.4	-1.0	4.4
49	-0.7	-1.1	0.4
50	-0.3	0.2	0.5
51	-0.2	-0.9	0.7
52	-0.2	-0.3	0.1
53	-0.2	-0.1	0.1
54	-0.2	-0.2	0.0
55	-0.3	-1.2	0.9
56	-0.6	-0.2	0.4
57	-0.1	0.6	0.7
58	-0.6	2.2	2.8
59	0.6	0.4	0.2
60	-0.4	0.0	0.4
61	0.2	-0.0	0.2
62	0.6	-0.4	1.0
63	-1.3	-1.0	0.3
64	-0.2	-0.5	0.3
65	1.7	0.2	1.5
66	-0.1	0.1	0.2
67	-0.1	-0.2	0.1

fit the Rasch model (Fit-t = -1.0). Item 48 in the Fall v. Spring is shown in Figure 5.

Item 48, Fall: Gerry's government job is a GS-12. Is GS-12:

A. Nominal
B. Ordinal
C. Interval
D. Ratio

Item 48, Spring: Sylvia won the silver medal in speed skating. Is silver medal:

A. Nominal
B. Ordinal
C. Interval
D. Ratio

Figure 5.--Illustration of the revision of Item 48.

Clearly, the Fall version of Item 48 measured application of knowledge relating to a second variable--namely, knowledge about government job classifications! The Spring version of Item 48, administered just after the 1984 Winter Olympic Games, more directly measured knowledge of the scales of measurement, since everyone would presumably be familiar with the meaning of a silver medal! Item 48 concurrently "became easier" in Spring, since it was more directly measuring measurement concepts. This finding posed an interesting characteristic of application items, since it would appear that the novel situation presented in the item must be sufficiently familiar to the persons taking the item.

Other items which varied in difficulty provided feedback for instructional revision. For example, the formula necessary to solve Item 67 was given in Fall, but not in Spring. Thus, the taxonomic classification of the item changed from application to analysis and, concurrently made the item more difficult in the Spring. This was an instructional oversight that was not detected before the item analyses!

Thus, it was clearly advisable to revise a "faulty item", or to examine instructional practices rather than to simply delete an item from the item bank.

Item Validity

In order to investigate the validity of each item, the Rasch item difficulty of each item was regressed on the item rating (C_I). The results from this analysis are reported in Table 3. The overall correlation between observed v. expected item difficulty, which was operationally defined as item validity, was $r = -0.47$ ($p < .01$). However, inspection of Table 3 would reveal that some of the individual item validity predictions were errant. As with the item reliability analysis, "faulty" items were, in general, misunderstood in the item construction process. These items provided feedback for revision of either the item or the instruction.

The validity of each item was evaluated by the standardized residual from the regression of observed Rasch item difficulty on item rating. From the "t" column of Table 3, it can be seen that several individual items were of questionable validity. Any item whose standardized residual exceeded $|1.0|$ was adjusted by adding or subtracting 1 from the original C_I. These revised item ratings are reported in the column labelled "C_I'." The C_I' item ratings would be used for the item bank. The correlation between observed Rasch item difficulty and C_I' was $r = -0.67$. Approximately 90 percent of the reliable variance in the test items could be explained from the C_I' estimates.

The apparent overall validity of the test item difficulties would support the use of the criterion-referenced grading procedure (see Methods section). The full summary of this procedure is reported in Table 4. The criterion score range for each letter grade reported at the bottom of Table 4 has been demonstrated to be reliable and valid.

Table 3

Results of Item Validity Analysis

Item No.	Rating C_I	Obs. d_s	Pred. d_s	t	Rating C_I'	Item No.	Rating C_I	Obs. d_s	Pred. d_s	t	Rating C_I'
01	10	-0.31	-1.53	1.2	9	41	6	0.20	-0.37	0.6	6
02	4	2.09	0.21	1.9	3	42	8	-1.11	-0.95	-0.2	8
03	5	-0.53	-0.08	-0.4	5	43	6	0.48	-0.37	0.9	6
04	4	-1.55	0.21	-1.8	5	44	3	2.25	0.50	1.8	2
05	3	0.83	0.50	0.3	3	45	6	1.76	-0.37	2.1	5
06	4	-1.11	0.21	-1.3	5	46	8	-2.27	-0.95	-1.3	9
07	3	0.05	0.50	-0.4	3	47	5	1.03	-0.08	1.1	4
08	4	-0.78	0.21	-1.0	5	48	1	0.93	1.08	-0.1	1
09	4	-0.53	0.21	-0.7	4	49	9	0.48	-1.24	1.7	8
10	4	1.13	0.21	0.9	4	50	8	-1.55	-0.95	-0.6	8
11	2	1.23	0.79	0.4	2	51	4	0.20	0.21	-0.0	4
12	2	-0.05	0.79	-0.7	2	52	4	-0.53	0.21	-0.7	4
13	2	0.48	0.79	-0.3	2	53	5	-1.11	-0.08	-1.0	6
14	7	-0.78	-0.66	-0.1	7	54	6	0.05	-0.37	0.4	6
15	5	-0.78	-0.08	-0.7	5	55	3	0.60	0.50	0.1	3
16	6	-1.11	-0.37	-0.7	6	56	3	1.67	0.50	1.2	2
17	4	0.34	0.21	0.1	4	57	4	-0.31	0.21	-0.5	4
18	5	-1.11	-0.08	-1.0	6	58	4	1.32	0.21	1.1	3
19	2	0.05	0.79	-0.7	2	59	4	0.72	0.21	0.5	4
20	2	-1.11	0.79	-1.9	3	60	5	0.48	-0.08	0.6	5
21	4	-1.55	0.21	-1.8	5	61	4	0.20	0.21	-0.0	4
22	6	-1.11	-0.37	-0.7	6	62	4	1.67	0.21	1.5	3
23	8	-2.27	-0.95	-1.3	9	63	4	0.93	0.21	0.7	4
24	3	-0.53	0.50	-1.0	4	64	8	-0.31	-0.95	0.6	8
25	4	-0.78	0.21	-1.0	5	65	2	0.72	0.79	-0.1	2
26	7	-2.27	-0.66	-1.6	8	66	4	-0.31	0.21	-0.5	4
27	4	1.13	0.21	0.9	4	67	2	2.75	0.79	2.0	1
28	4	1.13	0.21	0.9	4						
29	4	1.59	0.21	1.4	3						
30	4	-0.12	0.21	-0.3	4						
31	6	0.34	-0.37	0.7	6						
32	6	-2.27	-0.37	-1.9	7						
33	6	-0.12	-0.37	0.3	6						
34	6	-1.11	-0.37	-0.7	6						
35	6	0.20	-0.37	0.6	6						
36	3	0.72	0.50	0.2	3						
37	7	-2.27	-0.66	-1.6	8						
38	6	1.13	-0.37	1.5	5						
39	6	-0.53	-0.37	-0.2	6						
40	4	1.13	0.21	0.9	4						

$\bar{C}_I' = 4.75$

$SD_{C_I'} = 1.98$

$r_{d_s C_I'} = -0.68$

Table 4

Determination of Criterion Scores
for Assignment of Letter Grades

Item No.	No. Options	Rating C_I	P S	A+ 1	A 2	A- 3	B+ 4	B 5	B- 6	C+ 7	C 8	C- 9	F 10
01	4	10	89	1	1	1	1	1	1	1	.90	.90	.80
02	4	4	51	1	.90	.90	.80	.75	.50	.50	.25	.25	.25
03	4	5	91	1	1	.90	.90	.80	.75	.50	.50	.25	.25
04	4	4	96	1	.90	.90	.80	.75	.50	.50	.25	.25	.25
05	4	3	75	.90	.90	.80	.75	.50	.50	.25	.25	.25	.25
06	4	4	95	1	.90	.90	.80	.75	.50	.50	.25	.25	.25
07	4	3	86	.90	.90	.80	.75	.50	.50	.25	.25	.25	.25
08	5	4	93	1	.90	.90	.80	.80	.60	.40	.20	.20	.20
09	5	4	91	1	.90	.90	.80	.80	.60	.40	.20	.20	.20
10	5	4	70	1	.90	.90	.80	.80	.60	.40	.20	.20	.20
11	5	2	68	.90	.80	.80	.60	.40	.20	.20	.20	.20	.20
12	5	2	86	.90	.80	.80	.60	.40	.20	.20	.20	.20	.20
13	5	2	81	.90	.80	.80	.60	.40	.20	.20	.20	.20	.20
14	4	7	93	1	1	1	1	.90	.90	.80	.75	.50	.50
15	4	5	93	1	1	.90	.90	.80	.75	.50	.25	.25	.25
16	4	6	95	1	1	1	.90	.90	.80	.75	.50	.50	.50
17	4	4	82	1	.90	.90	.80	.75	.50	.50	.25	.25	.25
18	4	5	95	1	1	.90	.90	.80	.75	.50	.50	.50	.25
19	4	2	86	.90	.80	.75	.50	.50	.25	.25	.25	.25	.25
20	4	2	95	.90	.80	.75	.50	.50	.25	.25	.25	.25	.25
21	5	4	96	1	.90	.90	.80	.80	.60	.40	.20	.20	.20
22	5	6	95	1	1	1	.90	.90	.80	.80	.60	.40	.20
23	5	8	98	1	1	1	1	1	.90	.90	.80	.60	.40
24	5	3	91	.90	.90	.80	.80	.60	.40	.20	.20	.20	.20
25	5	4	93	1	.90	.90	.80	.80	.60	.40	.20	.20	.20
26	5	7	98	1	1	1	1	.90	.90	.80	.80	.60	.40
27	5	4	70	1	.90	.90	.80	.80	.60	.40	.20	.20	.20
28	5	4	70	1	.90	.90	.80	.80	.60	.40	.20	.20	.20
29	5	4	61	1	.90	.90	.80	.80	.60	.40	.20	.20	.20
30	5	4	88	1	.90	.90	.80	.80	.60	.40	.20	.20	.20
31	5	6	82	1	1	1	.90	.90	.80	.80	.60	.40	.20
32	5	6	98	1	1	1	.90	.90	.80	.80	.60	.40	.20
33	5	6	88	1	1	1	.90	.90	.80	.80	.60	.40	.20
34	4	6	95	1	1	1	.90	.90	.80	.75	.50	.25	.50
35	4	6	84	1	1	1	.90	.90	.80	.75	.50	.25	.25
36	4	3	77	.90	.90	.80	.75	.50	.50	.25	.25	.25	.25
37	4	7	98	1	1	1	1	.90	.90	.80	.75	.50	.50
38	4	6	70	1	1	1	.90	.90	.80	.75	.50	.50	.25
39	4	6	91	1	1	1	.90	.90	.80	.75	.50	.50	.25
40	4	4	70	1	.90	.90	.80	.75	.50	.50	.25	.25	.25

TABLE 4, continued

Item No.	No. Options	Rating C_I	P S	A+ 1	A 2	A- 3	B+ 4	B 5	B- 6	C+ 7	C 8	C- 9	F 10
41	4	6	84	1	1	1	.90	.90	.80	.75	.50	.50	.25
42	4	8	95	1	1	1	1	1	.90	.90	.80	.75	.50
43	4	6	81	1	1	1	.90	.90	.80	.75	.50	.50	.50
44	4	3	47	.90	.90	.80	.75	.50	.50	.25	.25	.25	.25
45	4	6	58	1	1	1	.90	.90	.80	.75	.50	.50	.25
46	4	8	98	1	1	1	1	1	.90	.90	.80	.75	.50
47	4	5	72	1	1	.90	.90	.80	.75	.50	.50	.25	.25
48	4	1	74	.80	.75	.50	.50	.25	.25	.25	.25	.25	.25
49	4	9	81	1	1	1	1	1	1	.90	.90	.80	.75
50	4	8	96	1	1	1	1	1	.90	.90	.80	.75	.50
51	4	4	84	1	.90	.90	.80	.75	.50	.50	.25	.25	.25
52	4	4	91	1	.90	.90	.80	.75	.50	.50	.25	.25	.25
53	4	5	95	1	1	.90	.90	.80	.75	.50	.50	.25	.25
54	4	6	86	1	1	1	.90	.90	.80	.75	.50	.50	.25
55	4	3	79	.90	.90	.80	.75	.50	.50	.25	.25	.25	.25
56	4	3	60	.90	.90	.80	.75	.50	.50	.25	.25	.25	.25
57	4	4	89	1	.90	.90	.80	.75	.50	.50	.25	.25	.25
58	4	4	67	1	.90	.90	.80	.75	.50	.50	.25	.25	.25
59	4	4	77	1	.90	.90	.80	.75	.50	.50	.25	.25	.25
60	4	5	81	1	1	.90	.90	.80	.75	.50	.50	.25	.25
61	4	4	84	1	.90	.90	.80	.75	.50	.50	.25	.25	.25
62	4	4	60	1	.90	.90	.80	.75	.50	.50	.25	.25	.25
63	5	4	74	1	.90	.90	.80	.80	.60	.40	.20	.20	.20
64	5	8	89	1	1	1	1	1	.90	.90	.80	.60	.40
65	5	2	77	.90	.80	.80	.60	.40	.20	.20	.20	.20	.20
66	4	4	89	1	.90	.90	.80	.75	.50	.50	.25	.25	.25
67	4	2	37	.90	.80	.75	.50	.50	.25	.25	.25	.25	.25

Grade	Sum of Success Probabilities[3]	Range of Total Raw Score	Range of Latent Ability
A+	65.40	63 - 67[1]	above 3.25[2]
A	62.38	60 - 62	2.59 to 3.24
A-	60.65	57 - 59	2.13 to 2.58
B+	55.20	52 - 56	1.55 to 2.12
B	50.40	48 - 51	1.16 to 1.54
B-	42.30	40 - 47	0.50 to 1.15
C+	33.50	31 - 39	-0.18 to 0.49
C	25.50	23 - 30	-0.81 to -0.17
C-	21.60	19 - 22	-1.16 to -0.82
F	———	0 - 18	below -1.17[2]

[1]Note that the lowest score in the interval corresponds to one SE_M below the Sum of Success Probabilities (SE_M = 2.88)

[2]Perfect and zero total raw scores cannot be estimated using Rasch ability.

[3]Note that, once banked, the response probabilities can be estimated directly from the Rasch model. It would, however, be advisable to set the minimum probability at chance level. Also, SE_M could be estimated at each level

Relationship Between Item Reliability and Item Validity

The results from the analyses of item reliability and item validity pointed to a logical relationship between these two essential features of an item. In order to empirically test the statistical significance of this apparent relationship, each item was dichotomized as being either of acceptable or unacceptable item reliability and/or item validity. The resulting 2 x 2 contingency table was analyzed to estimate the Phi-coefficient, ϕ. Since the cells varied considerably from an even p v. q split (i.e. most items were of acceptable reliability and validity), the ϕ was adjusted by computing ϕ_{obs} / ϕ_{max}. The resulting $\phi = 0.33$ was statistically significant beyond $p < .01$, suggesting that items tending to be of poor reliability also tended to be of poor validity. Thus, modifying items or instructional practices should result in a useful item bank containing items of high reliability and validity.

Establishing the Item Bank

Final estimates of item difficulty and C_I are reported in Table 1. Item bank difficulty (d_B) and standard error ($Se(d_B)$) are shown along with the item rating, C_I', used for criterion-referenced grading. Item difficulty estimates were established by the average of d_F and d_S; standard errors were established by averaging $Se(d_F)$ and $Se(d_S)$. Item ratings were established by procedures discussed previously.

The resulting item bank for the midterm examination was subsequently put into a computer file and will serve as the basis for assembling future midterm examinations. Beginning in Fall Semester, 1984, new (i.e. pilot) items will be used in combination with banked items so that the item bank can be expanded. The process is, obviously, a cumulative one-- eventually resulting in a large item bank. Thus, it appeared feasible to achieve the major goal of this study-- to develop a useful item bank for a teacher-made test!

Limitations

The procedures for developing the item bank were clearly arduous! However, given refinement of the details used for this study, the item banking procedures were clearly shown to be encouraging for further development. In particular, the need for developing computer software to facilitate the banking procedures would be tremendously helpful. Given the development of appropriate computer software, it would be quite feasible for a testing organization to offer item banking services to the educational community.

CONCLUSIONS AND IMPLICATIONS

The purpose of this study was to examine the feasibility of creating an item bank from a teacher-made test administered to two classes of comparable students instructed by the same teacher using the same lecture notes. In order to investigate this problem, two graduate Introduction to Measurement classes were given the same, 67-item midterm examination developed by the course instructor, who taught the course during Fall Semester, 1983 and Spring Semester, 1984. The feasibility of banking the test items would depend upon the comparability of two essential qualities of each item--invariant item difficulty (i.e. item reliability) and item quality (i.e. item validity). For the analysis, a Rasch latent trait analysis was performed on each of the two test administrations. The reliability of each item difficulty from Fall v. Spring was investigated by comparing the difference between the two Rasch difficulty estimates for each item against their calibration error. Evidence of item validity was studied by examining the extent to which each item approximated an a priori rating used for criterion-referenced grading.

The results indicated that the majority of items could be usefully banked based upon evident reliability and validity. Therefore, the prospect

of banking teacher-made test items appears favorable, given comparable instructional conditions and careful item editing procedures. The important implication of this study is that assembling subsequent examinations containing "different items" with known, reliable item difficulties and high validity indices, would be feasible. Thus, even teacher-made tests could be designed to optimally measure student achievement without being bound to re-use the same set of items. The item banking procedure would, therefore, facilitate high internal consistency reliability and construct validity for teacher-made tests.

REFERENCES

Berk, R.A. (ed.). Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press, 1980.

Bloom, B.S., Hastings, J.T., and Madaus, G.F. Handbook on formative and summative evaluation of student learning. New York: McGraw Hill, 1971.

Choppin, B. An item banking procedure using sample-free calibration. In Wood, R. and Skurnik, L.S. Item banking. National Foundation for Educational Research in England and Wales, 1969.

Ebel, R.L. The practical validation of tests of ability. Educational Measurement: Issues and Practice, 1983, 3, 7-10.

Glass, G.V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-62.

Loevinger, J. Objective tests as instruments of psychological theory. Psychological Reports, 1957, 3, 635-94.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Thorndike, R.L. and Hagen E.P. Measurement and evaluation in psychology and education (4ed). New York: Wiley, 1977.

Wright, B.D. and Mead, R.J. BICAL: Calibrating items and scales with the Rasch model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, The University of Chicago, 1976.

Wright, B.D. and Stone, M.H. Best test design. Chicago: MESA Press, 1979.